Title: Determining prescriptions in electronic health care (EHR) data: methods for development of standardised, reproducible drug codelists

Authors: Emily L Graul[1], Philip W Stone[1,2], Georgie M Massen[2], Sara Hatam[3], Alexander Adamson[1,2], Spiros Denaxas[4,5], Nicholas S Peters[2], Jennifer K Quint[1,2]

Addresses:

1. School of Public Health, Imperial College London, London, UK
2. National Heart & Lung Institute, Imperial College London, London, UK
3. Usher Institute, University of Edinburgh, Edinburgh, UK
4. University College London, Institute of Health Informatics, London, UK
5. British Heart Foundation Data Science Centre, Health Data Research UK, London, UK

Corresponding Author: Emily L Graul

School of Public Health, Imperial College London – White City Campus

86 Wood Lane

London, UK W12 0BZ

Phone: 02075948821

Email: e.graul@imperial.ac.uk

Key words: Codelist, Prescriptions, Value Sets, Electronic Health Records, epidemiology

Word Count: 3919/4000

ABSTRACT

Objective

Epidemiological research using electronic healthcare records(EHR) uses combinations of codes to define diseases and prescriptions (or phenotypes), requiring transparency and reproducibility. Yet methodology to generate codelists varies, manifesting in misclassification bias. Therefore we designed methodology enabling codelist reproducibility and generalisability across study contexts.

Materials and Methods

We developed a process to generate drug codelists, testing this using the Clinical Practice Research Datalink (CPRD) Aurum database, accounting for missing data in 'attribute' variables searched to identify codes. We generated a 1) cardiovascular codelist and 2) codelist for inhaled Chronic Obstructive Pulmonary Disease (COPD) therapies, applying them to a sample cohort of 335,931 COPD patients. We compared searching on all search variables (A, "gold standard") to B) chemical and C) ontological information only.

Results

In our full search (A), within follow-up, we determined 165,150 patients (49.2% of cohort) prescribed drugs from the cardiovascular codelist. For the COPD inhalers codelist, we determined 317,963 patients (94.7% of cohort) prescribed. Considering output within each individual value set, Search C missed substantial prescriptions, including vasodilator anti-hypertensives (19,696 prescriptions for A and B; 1,145 for C), and for SAMA (35,310 for A and B; 564 for C).

Discussion

Regardless of database and study context, we recommend the full method (A) for comprehensiveness. Despite database used, there are special considerations when generating adaptable drug codelists, including fluctuating status, cohort-specific drug indications, underlying hierarchical ontology, and collinearity in covariate analyses.

Conclusions

Generating drug codelists must use standardisable and reproducible methodology based on underlying ontology, with end-to-end clinical input.

BACKGROUND AND SIGNIFICANCE

Health data research & codelist generation

Research using electronic health records (EHR) is increasingly used to inform patient care across a breadth of longitudinal data sources, including the U.S. Veterans EHR System, the INSIGHT Clinical Research Network (CRN) database, the National Patient-Centered Clinical Research Network (PCORnet), the National COVID Cohort Collaborative (N3C) Data Enclave, Duke Health Data Warehouse, the SAIL Databank, the Clinical Practice Research Datalink (CPRD), NHSDigital and Northwest London Whole Systems Integrated Care (WSIC), with some allowing linkage to mortality, socioeconomic, registry and audit data[1–11].

Determining exposures, outcomes and covariates is central to EHR research[12,13] through generation of medical and drug "codelists" for an overarching clinical definition. Unfortunately, both methodology and reporting vary, forming sources of misclassification bias when ascertaining conditions and prescriptions. Methodology may not involve clinician review, manifesting in exclusion of necessary codes alongside inclusion of inappropriate codes. The Reporting of studies Conducted using Observational Routinely-collected health Data (RECORD) statement calls for EHR studies to provide "a complete list of codes and algorithms used to classify exposures, outcomes, confounders, and effect modifiers…considering the risk of misclassification bias…authors should provide sufficient detail to make their research *reproducible* and to make the *risk of bias* apparent" (emphasis added)[14].

Calls regarding transparency and bias extend beyond making codelists freely accessible in repositories. It also requires a systematic and standardisable yet malleable approach incorporating clinical input, designed around proactively considering when bias can manifest upon subsequent application to cohorts. Malleability within codelist design can make reproducibility and generalisation to other contexts and databases possible, allowing consistent definition (harmonisation) of phenotypes[15]. Malleability facilitates researchers to adapt and reuse others' codelists for their study and, conversely, facilitates researchers to contribute their codelists to others' studies appropriately[16]. Literature on drug codelists has been primarily high-level[17,18], underlying steps for its generation of the codelists themselves not usually described.

Aim

Given unique considerations for drug codelists, we developed a standardisable, reproducible method for creating drug codelists that incorporates end-to-end clinicial expertise, considers missing data in search 'attribute' variables, considers fluctuating status, and is adaptable to other studies and databases. We then operationalise the methodology to generate a codelist for oral drugs for hypertension and heart failure, and a disease-specific codelist, all inhaled therapies for COPD. We also apply the codelists to a sample cohort, according to study considerations and clinical review.

MATERIALS AND METHODS

Defining phenotypes, value sets, and ontologies

Common to both medical and drug codelist generation is establishing the single overarching clinical definition (analogous to a "phenotype" for medical codelists), premised by a database's underlying drug hierarchical ontology (e.g., British National Formulary, BNF, Anatomic Classification System, ATC, and RxNorm)[19–21]. Within this definition are "value sets", a "uniquely identifiable set of valid concept representations" for "possible values of a coded data element in an information model"[22]. Depending on purpose and study context there may be one or multiple value sets. Ideally, searches to identify all possible codes are conducted purely using the drugs' chemical terms or the database's ontology, but missing data can prevent this.

Our methodology

Our methodology for generating drug codelists has seven steps (summarised in Figure 1) and is available on our Github repository. It is based on our methodology for generating medical codelists, also available in our Github repository.

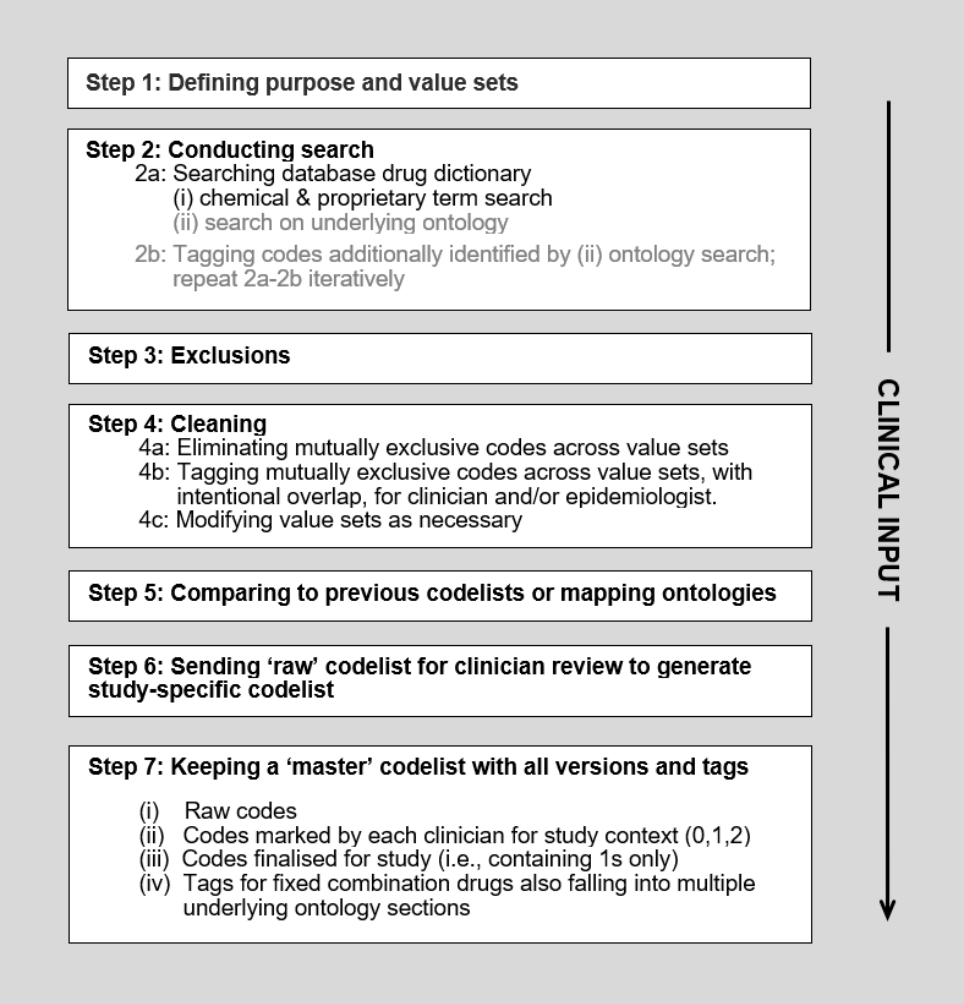

Figure 1. Summary of methodology for generating a product codelist
Note: Steps in grey font may be database-dependent and modified as necessary given missing data.

Step 1: Defining purpose and value sets

The first stage is to identify value sets consisting of drug class(es) based on an underlying ontology and within classes a list of drugs with corresponding chemical and potentially multiple proprietary names. The route(s) of administration (e.g., oral, parenteral, ocular) should be specified. The aim may be to produce a broader codelist permitting modification for various contexts, or to generate a study-specific codelist.

We consider clinical input regarding chemistry to improve search efficiency. For example, searching on common compounds, active or blocking groups, or side chains such as *-nitrate -arginine -hydrochloride -mesilate* is not recommended. Whilst these suffixes may be listed as part of the drug name, they are not the chemical-of-interest and may lead to inefficiently large search outputs.

Premising step 1 is confirming the database's underlying ontology to systematically collate a list of all possible terms (i.e., the BNF for UK databases) and reliable resource to facilitate reproducibility of collation. The most user-friendly BNF resource is OpenPrescribing [23], a interface utilizing raw data from UK National Health Service Business Services Authority (NHSBSA), with an ability to easily extract corresponding proprietary names and identify discontinued drugs.

Step 2: Conducting search

The second step is to search based on the collated list, using wildcard (*) characters to pick up terms in any location within a string.

Step 2a: Search the database drug dictionary

Before the search, we import the database's drug 'dictionary' text file. We import all 'attribute' variables searched upon as strings.

We search the browser in 2 stages:

(i) For each drug within value sets, search by the chemical and all proprietary names, for a string match on the associated 'attribute' variables.

(ii)Search on underlying ontology, considering syntax with slashes(/) as medicines may be indicated for multiple conditions and located in chapters not primarily of interest but still used for the indication-of-focus (e.g., searching for sildenafil (parenteral route) may fall into section ontologies for both hypertension and erectile dysfunction).

This automated search nests chemical and proprietary terms within each drug list, with corresponding lists nested within broader value sets (Figure 2), in effect sorting output for (i) by value set.

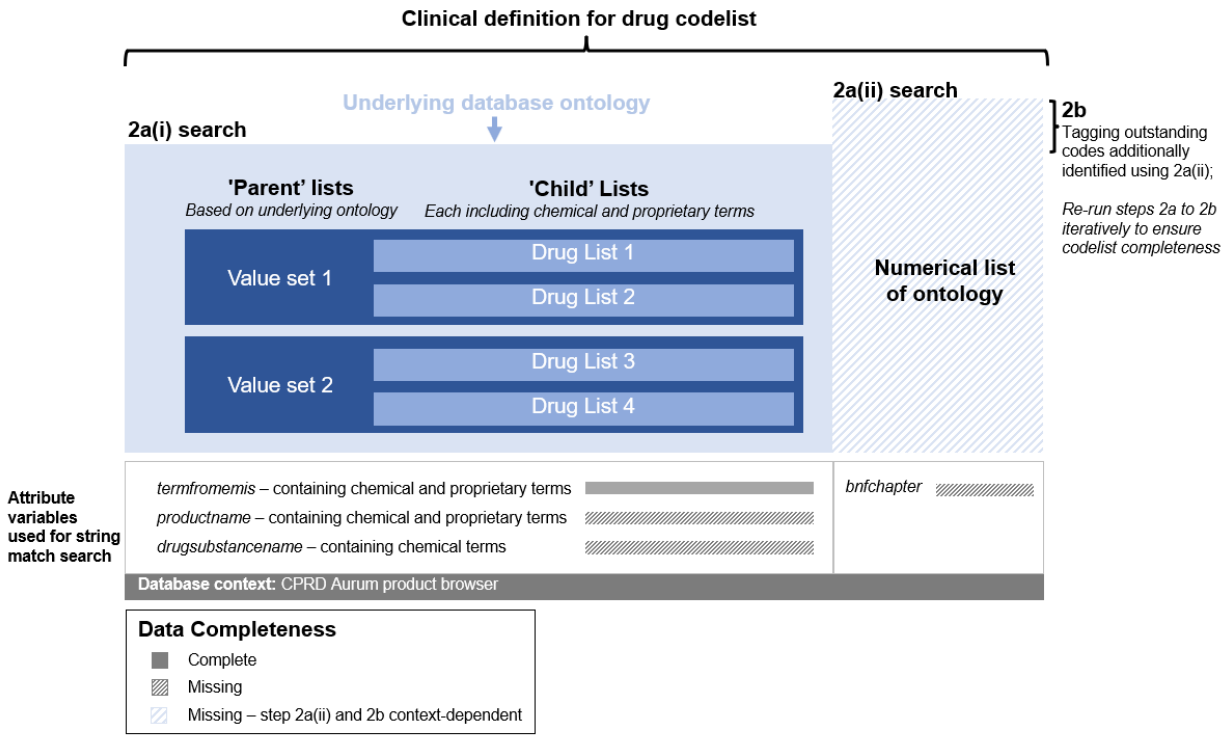

Figure 2. Flow diagram showing the Step 2 search process for drug codes.
A main clinical definition for the drug codelist is established, premised by the underlying database ontology. Within each given value set ('parent' list) are nested 'child' lists each corresponding to individual drugs, chemical and proprietary names. The CPRD-specific search attributes are *termfromemis* and *productname* (containing chemical and proprietary information) and *drugsubstancename* (chemical information). Due to missing data within the search 'attribute' variables we search on all three variables in Step 2a(i), with an additional ontological code search in Step 2a(ii), checking for search term completeness of the former by comparing the (i) and (ii) in Step 2b iteratively. Therefore, Step 2a(ii) and 2b may be database dependent given missing data. In Stata, parent and child lists take the form of local macros; in R a comparable step would be to name a list of vectors, and nesting the lists as necessary. The full Stata and R code including all drug codelist generation steps is located on our Github repository.

Step 2b: Tag codes additionally identified by searching on (ii) underlying ontology; Repeat 2a-2b iteratively

We tag outstanding codes not found by searching on chemical and proprietary terms alone, allowing researchers to check if they included all possible terms. Outstanding codes are identified and tagged if the row total for columns with "1" denoting searches (i) and (ii) within 2a sum to least two. If outstanding codes are present, one should add the additional names to the value sets, re-running steps 2a to 2b. Ideally, upon multiple iterations of re-running, there should be an absence of tags, indicating inclusion of all appropriate terms. This additional step may seem redundant but is most important to check codelist completeness.

In place of this step we incorporated an expanded search using Systematized Nomenclature of Medicine-Clinical Terms (SNOMED-CT, a commonly-used international reference terminology)

concept IDs to check for outsanding synonym codes fitting our value sets, as recommended for medical codelists[24]. However, in CPRD, although for clinical events, a given SNOMED-CT Concept ID will match with multiple SNOMED-CT description IDs, for prescription events, a given SNOMED-CT Concept ID does not match with more than one SNOMED-CT Description ID. This is because the UK SNOMED-CT drug extension is derived from the Dictionary of Medicines and Devices (dm+d), with the SNOMED-CT Concept ID being identical to the dm_d code, and CPRD only utilises dm+d codes, preventing this approach.

Step 3: Exclusions

Step three consists of, after manual review, excluding codes. This is distinguished from later exclusion based on epidemiological and clinical considerations specific to study context. Elimination may be based on information from drug name, route, and/or formulation. The broad search may pick up different medications with the same active chemical but of an inappropriate route, i.e., for a different medical indication corresponding to a different organ system.

We do not eliminate by product identifiers; it is a less transparent method. We do not eliminate by ontology chapter either, not only due to missing data, but anecdotal evidence suggests some drugs may have intended medical indication(s) corresponding to multiple chapters, which cannot be assumed and is not reflected in its ontological classification code.

Step 4: Cleaning

Step 4a: Eliminate mutually exclusive codes across value sets

This step places a temporary tag on codes mutually exclusive across value sets, a possibility given the broad search, e.g., even within a single BNF class chapter, active ingredients may overlap among the individual sections.

This tag allows researchers to write code automating the re-sorting process to make these sets exclusive.

205 ***Step 4b. Tagging mutually exclusive codes across value sets, with intentional overlap, for***
206 ***clinician and/or epidemiologist***
207
208 This step consists of proactively placing permanent tags on codes corresponding to fixed combination
209 drugs with 'intentional overlap.'
210
211 We define 'intentional overlap' when one code corresponds to a fixed combination drug consisting of
212 two drug classes (i.e., mechanisms of action) such that it resides in multiple ontological sections, e.g.,
213 hydrochlorothiazide/captopril, a single drug including both diuretic and Renin-angiotensin system
214 components.
215
216 Clinician input is considered to provide a list of possible suffixes for the tags, e.g., "*azide*" for
217 diuretics, or "*pril*" for angiotensin-converting enzyme (ACE) inhibitors and angiotensin receptor
218 blockers (ARBs).
219
220 ***Step 4c. Modifying value sets as necessary***
221 This step allows for value set modification. For example, combining sets into a broader value set upon
222 study context or computational considerations (e.g., Stata has macro character limits).
223
224 **Step 5. Comparing to previous lists or mapping ontologies**
225 This step is for merging together and comparing current and previous codelist versions, and to merge
226 and map codes labelled under different ontologies(e.g., ATC-BNF mapping). Comparison facilitates
227 correct categorization and possible identification of outstanding codes. Beyond completeness,
228 mapping allows harmonisation and reproducibility to other database contexts[15].
229
230 This leads to the 'raw' codelist that is not study-specific, and ready for adaptation to a cohort through
231 clinical review.
232
233 **Step 6. Sending 'raw' codelist for clinical review to generate study-specific codelist**
234 Clinician(s) review the 'raw' codelist and each code (observation) is labelled as the following
235 'certainty' categories:

0 = “clear exclusion”

1 = “certainty” to be included

2 = “uncertainty” if to include for future sensitivity analyses

Review by one clinician trained in epidemiology and familiar with using the database is necessary, *at least*. When placing certainty categories, clinicians may consider:

- *cohort-of-interest* specific to their expertise/organ system. If multiple systems are involved, i.e., multimorbidity, multiple clinicians are necessary
- *operationalisation of codes in clinical settings* and *patient behaviour or prescription commonality* within context
- *database characteristics* (e.g., drug with substantially low number of issues may have low frequency of prescription events when later applied to cohort)

An additional step of resolving discordances among ‘certainty’ categories is required given multiple clinicians. Step 6 adapts a publication’s methods on generating medical codelists[12] after which the codelist is finalised for study context.

Step 7. Keeping a ‘master’ codelist

Researchers should keep a ‘master’ codelist with all versions and tags for reasons pertaining to malleability: to allow adaptation for sensitivity analyses, and to allow generalisation and harmonisation to other and across study contexts.

It should contain:

(i) raw codes – all codes not study-specific, initially generated by epidemiologist (broad and all-encompassing) sorted by ontology

(ii) clinical input – columns per clinician (≥1 column with 0/1/2s)

(iii) study-specific codes – column for tailoring to context (1s ‘certainty’ only)

(iv) tags for individual codes for fixed combination drugs also classified within different ontology chapters

Operationalising our methodology

The methodology was applied to generate two codelists, 1) drugs for hypertension and heart failure (BNF Chapter 2.5) and 2) all inhaled therapies for COPD (BNF Ch. 3.1.1-.2; 3.1.4; 3.2) in CPRD Aurum[23].

We estimated the number of respective prescriptions among a cohort with COPD within the Clinical Practice Research Datalink (CPRD) Aurum in England. Many individuals with COPD require inhaled therapies for reduced lung function and have co-morbidities, including cardiovascular disease[25]. In this open cohort, patients were included and started follow up at latest of 1 January 2010 if they 1) were diagnosed with validated codes for COPD[26], 2) were at least 40 years, 3) had continuous 1 year GP registration, and 4) data was deemed 'acceptable' quality. Follow-up ended on the earliest of: 31 December 2019, the last collection date, date of patient death, or transfer out of GP practice.

We compared output of the following searches, for the codelist and upon application to the cohort:

A. Using our full 'gold-standard' methodology, searching on chemical and proprietary terms and BNF codes (chemical and proprietary names on *termfromemis*, *productname* variables, chemical names only for *drugsubstancename* due the nature of this variable, BNF codes for *bnfchapter*)
B. Using our methodology, but searching on chemical names only (within *drugsubstancename*)
C. Using our methodology, but searching on the BNF code only (within CPRD's *bnfchapter*)

Outcome measures were product codes, drug issues, and prescriptions in overall and by value set. Analyses were conducted using Stata v17 (Texas, USA). We wrote final scripts in Stata, translating into R v4.2.0. A summary of each codelist's purpose and operationalisation is in Supplementary Table S1. R and Stata scripts, as well as a full list of our terms for the value sets, are in located in our Github repository.

RESULTS

Generating raw codelists

We operationalised the methodology in CPRD Aurum to generate two codelists, a cardiovascular codelist for hypertension and heart failure medication and a codelist for inhaled COPD therapies.

298

299 We designed value sets around the codelist purpose (e.g., repository or disease-focused). We nested
300 terms corresponding to each drug within ‘child’ lists, then nesting each drug list into ‘parent’
301 lists(Figure 2). For example, for the indoramin child list (cardiovascular codelist), we searched for
302 chemical and proprietary terms, nesting this list within the BNF chapter 2.5.4 value set). In the COPD
303 inhalers codelist, because intentions were to distinguish disease-specific drugs, we separated value
304 by type even though chemical compositions overlapped, e.g., inhaled corticosteroids (ICS) vs. inhaled
305 corticosteroid-long-acting muscarinic antagonists (ICS-LAMA).

306

307 In the CPRD Aurum drug dictionary, a given unique identifier (‘*prodcodeid*’ variable) can include
308 missing data on its following ‘attributes’ including active chemical ingredients (‘*drugsubstancename’*),
309 ontology (‘*bnfchapter*’), and route (‘*routeofadministration*’). (A description of missing data in the CPRD
310 drug dictionary is in Supplementary Table S2.) Ideally, searches to identify all possible codes would
311 be conducted purely using chemical terms or the database’s ontology, but missing data prevented
312 this, while the most-complete *termfromemis* variable lists drugs by *either* chemical or proprietary
313 name. Therefore, in in Step 2a(i) we searched on multiple ‘attribute’ variables: *termfromemis*,
314 *productname* and *drugsubstancename* (Figure 2).

315

316 Prior to producing raw codelists, we excluded 26 codes from the cardiovascular codelist, and 206
317 codes from the COPD inhalers codelist. This was due to cases where composition was correct but
318 route was incorrect given indication (e.g., *cutaneous* minodoxil, *ocular* guanethidine monosulfate,
319 salbutamol *nebuliser solutions*), our string match inadvertently picked up codes for a different purpose
320 or distinct chemical compound (e.g., Glutenex prescription from searching “*tenex*”; apra-clonidine
321 from searching on “*clonidine*”), or the term was not part of value sets (e.g., ICS-salbutamol codes).

322

323 We tagged codes for fixed combination drugs also classified within other BNF chapters. For the
324 cardiovascular codelist, this comprised codes for Ch. 2.2 diuretics and 2.6 antianginal drugs, e.g.,
325 Lisinopril-Hydrochlorothiazide (diuretic and angiotensin-converting enzyme (ACE) inhibitor). For the
326 COPD inhalers codelist this comprised codes for Ch. 3.3, for Salbutamol-sodium cromoglycate.

327

After respective exclusions and tags, the raw cardiovascular codelist contained 601 codes of both oral and parenteral routes, and for inhaled COPD therapies, 259 codes. For the COPD inhalers codelist, after subsequently merging with a previous codelist mapped to NHS Business Services Authority TRUD ATC-BNF ontology mapping files, this led to a final count of 472 codes. Of these codes, 77 were new codes not in the previous codelist, and 13 outstanding codes were from the previous codelist not in the new codelist. Most new codes were ICS- or short acting beta-agonist (SABA)-based (Supplementary Table S5).

Clinical review

The first clinician, a respiratory consultant and epidemiologist, removed codes for drugs not part of value sets: for the cardiovascular codelist, nine for Selexipag and one for Sodium Nitroprusside from searching on BNF ontology(Step 2a,ii). There was concern such prescriptions were rare given few issues and it being less likely these infrequently-used, new, or discontinued drugs were prescribed in the COPD cohort. For the COPD inhalers codelist, we removed three codes (0s) consisting of an ambiguous term for route – “liquid” or “solution”– potentially corresponding to nebulised therapies. We retained and tagged one code for paediatrics.

For the cardiovascular codelist, 33 codes of the parenteral route were given 0s as this route is not typically prescribed in UK primary care, leaving oral medications. A second clinician, a cardiologist, reviewed the 1s, agreeing on all codes.

Final codelists

The final cardiovascular codelist consisted of 568 codes for oral medications (Figure 3;Supplmentary TableS3), including the 66 and 28 product codes tagged for overlap with Chapter 2.2 and 2.6. The value set with the greatest count was for drugs targeting the Renin-aldosterone system (Ch. 2.5.5, N=375), whereas 2.5.3 and 2.5.8 were the smallest (N=4, N=2 respectively).

The final COPD inhalers codelist consisted of 456 codes(Figure 3;Supplmentary TableS3). The largest value sets were for ICS and long-acting beta-agonists (ICS-LABA)(N=201, N=71, respectively).

359 **Applying codelists to find prescriptions**

360 We applied the codelists to a cohort of 335,931 patients diagnosed with COPD, according to study
361 population considerations and clinical input.

362

363 For the cardiovascular codelist, within the decade follow-up, 165,150 patients (49.2% of cohort) were
364 prescribed at least one of the drugs (Figure 3;Supplmentary TableS3). As in the case with count, the
365 value set for chapter 2.5.5 had the greatest number of patients prescribed (N=151,225, 45.0% of
366 cohort). Chapters 2.5.3 and 2.5.8 did not have prescriptions.

367

368 For the COPD inhalers codelist, we determined 317,963 patients (94.7% of cohort) prescribed at least
369 one of the drugs (Figure 4;Supplmentary TableS4). Counts and prescriptions followed different
370 patterns. Whilst ICS had greatest count (N=213 codes), SABA had the most prescriptions
371 (N=297,966; 88.7% of cohort).

372

373 **Comparing to restricted searches**

374 We compared output of our full 'gold-standard' searches (A) to two restricted searches still using our
375 methodology but searching on (B) chemical terms within *drugsubstancename* and (C) BNF ontology
376 within *bnfchapter*.

377

378 For the cardiovascular codelist, search B identified 505 codes and 155,678 patients prescribed
379 (46.3% of cohort) at least one of the drugs across value sets. Search C identified 267 codes and
380 150,669 patients (44.9% of cohort) prescribed at least one of the drugs across value sets (Figure 3;
381 Supplementary TableS3).

382

383 For the COPD inhalers codelist, search B identified 351 codes, and 317,957 patients prescribed (95%
384 of cohort) at least one of the drugs across value sets. Search C identified 185 codes and 315,749
385 patients (94% of cohort) prescribed at least one of the drugs across value sets (Figure 4;
386 Supplementary TableS4).

387

388 The percent increase in output from searching on BNF ontology only (C) to the gold-standard (Search
389 A) was the most pronounced (cardiovascular codelist: 113% and 9.6% increase in codes and

prescriptions, respectively; COPD codelist 147% and 0.7%). However, we observed marginal increases when comparing Search A to searching solely on chemical information (B) (cardiovascular codelist: 12.5%, 6.08% increase respectively; COPD codelist 29%, 0.0002% increase respectively).

Considering restricted searches by value set, there were absent or marginal increases in prescriptions in some value sets upon using Search A (close to 0%), but remarkable increase in others, particularly for C (up to 24802% BNF 2.5.4; up to 6161% for SAMA). Search B led to higher output compared with C, except for Ch. 2.5.5. Here, greater counts in B vs. C did not translate to greater prescriptions (N=343 counts, N=138,992 prescriptions for B; N=217 and N=150,117 for C).

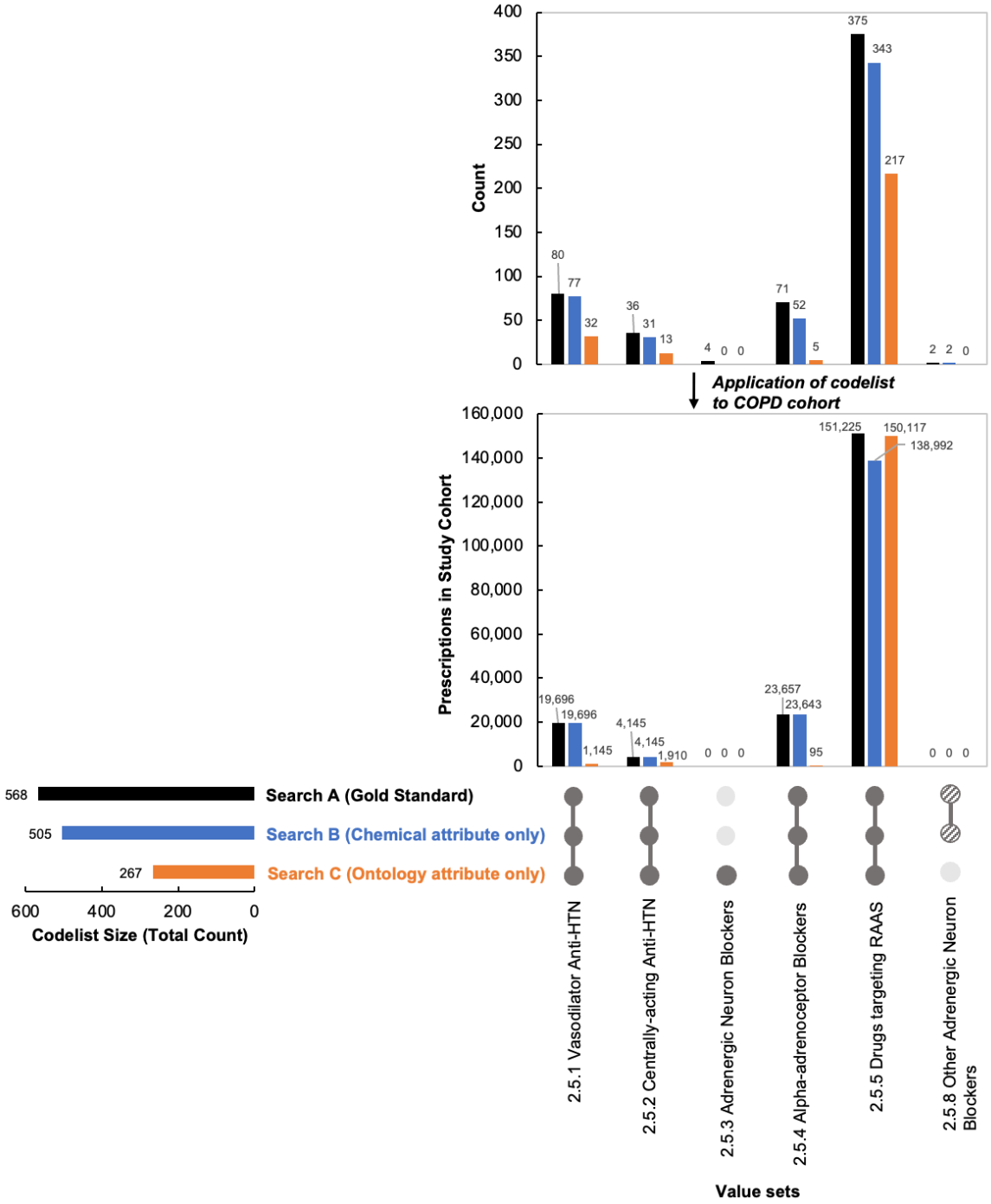

Figure 3. Comparison of codelist output by search type for the cardiovascular codelist
Adapted from an UpSet plot design[27,28]. RAAS=renin-angiotensin-aldosterone system. Results refer to post-clinicians' input. Counts derive from codelist generation, prescribed patients determined after codelist applied to COPD cohort. Prescriptions by value set are mutually exclusive as some patients are prescribed drugs of different classes across value sets, e.g., prescribed drug falling into 'at least' one subsection. Partially shaded dots for Ch. 2.5.8 refer to presence of codes in codelist, but absence upon application to cohort to determine prescriptions. Search A refers to the use of our methodology, searching *termfromemis, productname, drugsubstancename*, and *bnfchapter* variables. Search B refers to the use of our methodology, but searching on *drugsubstancename* variable, only. Search C refers to the use of our methodology, but searching on *bnfchapter* variable, only (Search C). Refer to Supplementary Table S3 for full data.

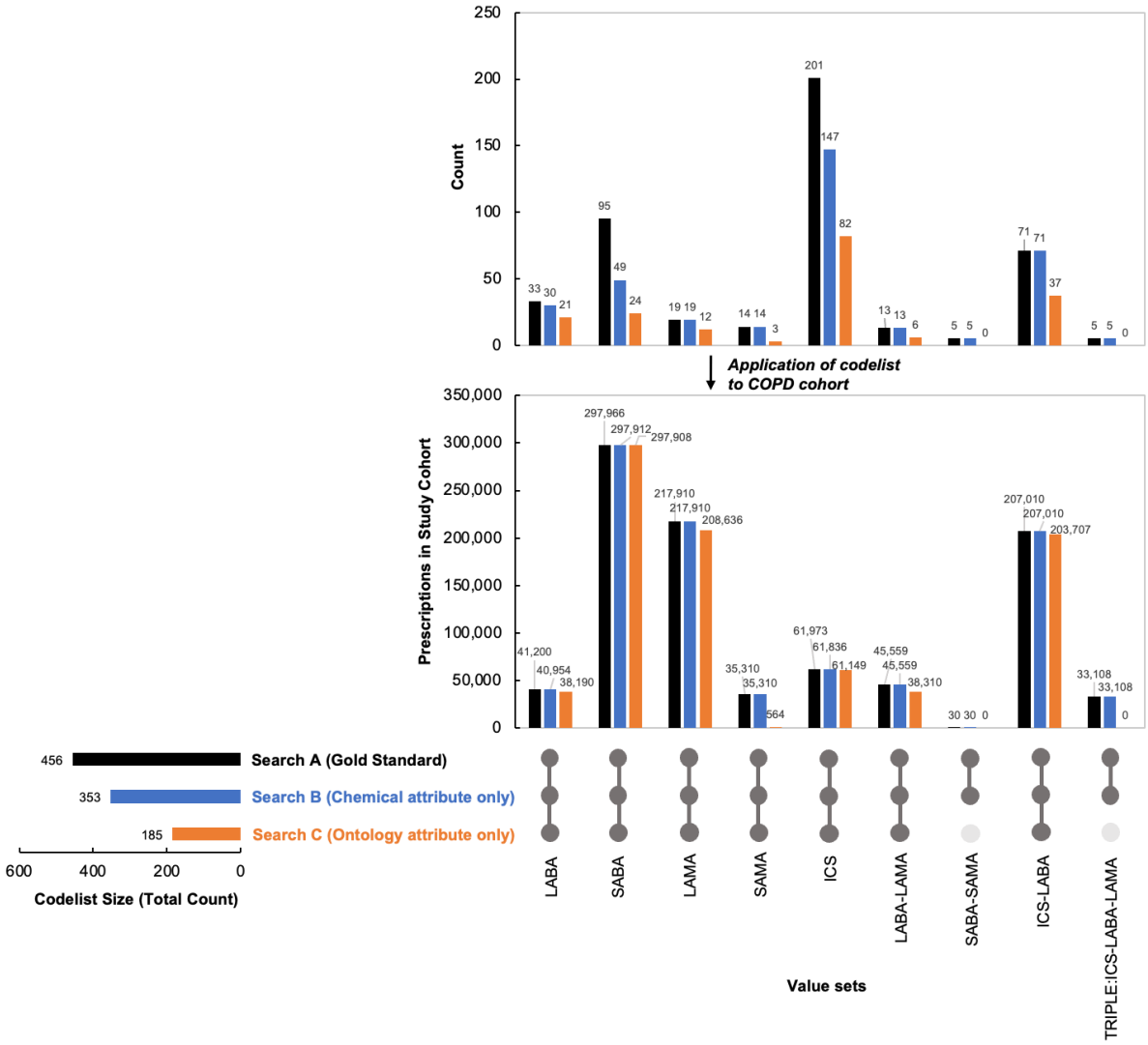

Figure 4. Comparison of codelist output by search type for the COPD inhalers codelist
Adapted from an UpSet plot design[27,28]. Results refer to post-clinicians' input. Counts derive from codelist generation, prescribed patients determined after codelist applied to COPD cohort. Prescriptions by value set are mutually exclusive as some patients are prescribed drugs of different classes across value sets, e.g., prescribed drug falling into 'at least' one subsection. Search A refers to the use of our methodology, searching *termfromemis, productname, drugsubstancename*, and *bnfchapter* variables. Search B refers to the use of our methodology, but searching on *drugsubstancename* variable, only. Search C refers to the use of our methodology, but searching on *bnfchapter* variable, only (Search C). Refer to Supplementary Table S4 for full data.

DISCUSSION

Summary

We developed a standardisable, reproducible method for creating comprehensive drug codelists using a semi-automated process incorporating end-to-end clinician expertise, considering missing data and fluctuating status that is adaptable to other studies and databases.

We applied the methodology to generate two codelists that were implemented on a sample cohort of patients with COPD in CPRD Aurum, according to study considerations and clinical review.

Evaluation

Searching only on BNF ontology (C) was insufficient and missed the most prescriptions in CPRD. Given similar results for searches A and B, there may be a marginal opportunity cost between searching on all variable attributes versus chemical terms alone, depending on drugs desired (i.e., restricting to a portion of the value sets, e.g., a short-acting muscarinic antagonist (SAMA)-only codelist; or drugs in 2.5.4 only), and cohort size (e.g., rare vs. common disease). Value set Ch. 2.5.5 results could be explained by, despite fewer counts, saturation of prescriptions among particular codes, upon codelist application to the study cohort.

Although we focus on CPRD and the cohort-of-interest, methods centre around addition of new information (e.g., drugs, proprietary names), completeness, and context-specific adaptations, applicable to other databases and underlying ontologies.

Recommendations

We recommend the full, ‘gold-standard’ method (A) if comprehensiveness and reaching statistical power (i.e. studies where codelist defines cohort or exposure) or when sample size (e.g., propensity scoring on complete data) is of focus. . In our database context which contained missing data in the drug dictionary, a comprehensive search on multiple ‘attribute’ variables was warranted, but this may not be the case for other data sources with data completeness.

If the aim is to produce a broader codelist (i.e., higher specificity, lower sensitivity) permitting modification for various contexts, we recommend limiting *a priori* exclusion criteria in step 1; rather, designing sets around underlying taxonomy. If the aim is disease-specific, value set generation should be clinician-led but still designed to permit malleability for different studies, i.e., single/fewer classes (e.g., statins only). To ensure all possible terms are found we recommend steps 2a-2b iteratively.

460 Comparisons to previous codelists and/or mapping files can pick up additional codes. In studies
461 incorporating multiple databases, we recommend merging with ontology mapping files to enable
462 codelist harmonisation[15,29], such as ATC-BNF mapping.
463
464 **Current literature**
465 Complexities of creating medical codelists are described elsewhere[12,24], and in some databases,
466 recorded prescriptions are treated separately from medical events and often recorded in separate
467 tables using different coding schemes[2,5,9]. Few studies outline methods for codelist development,
468 with the paucity of literature prominent for drug codelists where focus has been high-level, covering
469 challenges, assumptions, and principles in EHR research, e.g., software and analytical techniques,
470 data preparation, and defining periods for drug covariates[17,18], but not on underlying steps for
471 codelist generation. Of literature on medical codelists, the focus was on general guidelines for
472 researchers to follow to enhance reproducibility[12,30], including incorporating clinician review to
473 explore codes' uncertainty[12], exploring applications of codelists to sample cohorts[12], and
474 identifying and comparing disease phenotypes based on restricted versus expanded conceptual
475 definitions[16,24].
476
477 **Other considerations**
478 We recommend consideration of the database and context in all study stages, yet emphasize building
479 modifiability into codelist methodology to allow generalisation. Modifications may derive from study
480 *nature*, including the period (e.g., retrospective with discontinued drugs in-use during the study),
481 cohort-of-interest (e.g., patients with COPD where certain cardiovascular prescriptions are not
482 indicated), and subsequent statistical analyses. In studies with drug covariates, overlaps in class
483 could present collinearity, where exclusion of overlapping codes may be required. Our solution was
484 proactively tagging overlapping codes, drawing upon clinical expertise. Codelist tailoring may relate to
485 *database* factors, such as data type (e.g., medications recorded in primary care versus claims data)
486 and taxonomy (e.g., BNF or ATC codes).
487
488 Due to changes in new, existing, and discontinued drugs, and periods-of-interest for retrospective
489 studies, the same codelist may need to be updated based on older or newer database versions.
490 Using nested lists allows for maintained organisation despite these realities. Future methods may

consider adding an extra column to the codelist showing "in use" status, although anecdotal evidence suggests applying codes for discontinued drugs to a newer cohort will not pick up prescriptions.

Results on number of patients prescribed may differ depending on the cohort. Before generating final study-specific codelists, clinical input for exclusion may differ by disease, e.g., drugs contraindicated in specific patients such the codelist would be adapted to exclude corresponding value set(s). Prescriptions may change if the codelist is applied to separate disease cohorts, where prescription commonality by disease varies (e.g., cohorts with COPD, diabetes, Chronic Kidney Disease, or multimorbidity).

Results on number of prescriptions may differ depending on how codelists are operationalised, i.e., as covariates or exposures, part of inclusion, or accounting factors such as duration and frequency[31] and for combination, open therapies (e.g., for SAMA-SABA beyond fixed therapies determined through codelist generation).

Regarding external overlap to other taxonomic sections, the act of tagging is information proactively sought and built into initial steps, allowing for streamlined adaptability, e.g., should covariate collinearity arise. When considering adjustment of confounding covariates, covariate collinearity could distort observed exposure-outcome effects. Resolving collinearity through codelist adaptation should include clinical input.

CONCLUSION

We designed a semi-automated process to generate drug codelists using standardisable and reproducible methodology. Despite database identity, there are special considerations when generating adaptable drug codelists, including in-flux status, cohort-specific drug indication and exclusions, database-specific underlying hierarchical ontology, and operationalisation relating to inclusion and covariate analysis. Regardless, many EHR researchers are not clinicians; supplemental input is necessary.

AUTHOR CONTRIBUTIONS

EG devised the protocol, sought ethics approval, and supervised the project, with PS advising on methodology. PS, GM, and SH provided methodological and analytical support. EG drafted the manuscript. JQ provided respiratory clinical input when generating the two cohort-specific codelists. NP provided clinical input when generating the cardiovascular cohort-specific codelist. AA translated the STATA script into R. All authors contributed content, reviewed and approved the final manuscript.

FUNDING

No funding is reported for this study. This research was supported by the NIHR Imperial Biomedical Research Centre (BRC).

CONFLICTS OF INTEREST

JQ has received grants from MRC, HDR UK, GSK, BI, asthma+lung UK, and AZ and personal fees for advisory board participation, consultancy or speaking fees from GlaxoSmithKline, Evidera, AstraZeneca, Insmed. NP has received funding from Imperial Health Charity, SD is supported by the BHF Data Science Centre led by HDR UK (grant SP/19/3/34678), BigData@Heart Consortium, funded by the Innovative Medicines Initiative-2 Joint Undertaking under grant agreement 116074, the NIHR Biomedical Research Centre at University College London Hospital NHS Trust (UCLH BRC), a BHF Accelerator Award (AA/18/6/24223), E) the CVD-COVID-UK/COVID-IMPACT consortium and the Multimorbidity Mechanism and Therapeutic Research Collaborative (MMTRC, grant number MR/V033867/1). PS reports grants from asthma+lung UK and Gilead. EG, GM, AA, and SH have nothing to disclose.

PATIENT CONSENT FOR PUBLICATION

Not applicable.

ACKNOWLEDGEMENTS

The data used are from CPRD obtained under license from the UK Medicines and Healthcare products Regulatory Agency. The data is provided by patients and collected by the National Health Service as part of their care and support.

PEER REVIEW

Externally peer reviewed.

548 **DATA AVAILABILITY STATEMENT**

549 Data may be obtained from a third party and are not publicly available. Data are available on request
550 from CPRD. CPRD data provision requires purchase of a license, and this license does not permit the
551 authors to make them publicly available to all.

552 **ETHICS APPROVAL**

553 CPRD has NHS Health Research Authority (HRA) Research Ethics Committee (REC) approval to
554 allow the collection and release of anonymised primary care data for observational research [NHS
555 HRA REC reference number: 05/MRE04/87]. Each year CPRD obtains Section 251 regulatory
556 support through the HRA Confidentiality Advisory Group (CAG), to enable patient identifiers, without
557 accompanying clinical data, to flow from CPRD contributing GP practices in England to NHS Digital,
558 for the purposes of data linkage [CAG reference number: 21/CAG/0008]. The protocol for this
559 research was approved by CPRD's Research Data Governance (RDG) Process (protocol number:
560 22_002515) and the approved protocol is available upon request. Linked pseudonymised data was
561 provided for this study by CPRD. Data is linked by NHS Digital, the statutory trusted third party for
562 linking data, using identifiable data held only by NHS Digital. Select general practices consent to this
563 process at a practice level with individual patients having the right to opt-out.

564

565 **REFERENCES**

566 1 Xu Z, Wang F, Adekkanattu P, *et al.* Subphenotyping depression using machine learning and
567 electronic health records. *Learn Health Syst* 2020;**4**:e10241. doi:10.1002/lrh2.10241

568 2 NCATS National COVID Cohort Collaborative (N3C) Data Enclave. COVID-19 Clinical Data
569 Warehouse Data Dictionary: Based on OMOP Common Data Model Specifications Version 5.3.
570 https://ncats.nih.gov/files/OMOP_CDM_COVID.pdf

571 3 Justice AC, Gordon KS, Romero J, *et al.* Polypharmacy-associated risk of hospitalisation among
572 people ageing with and without HIV: an observational study. *Lancet Healthy Longev*
573 2021;**2**:e639–50. doi:10.1016/S2666-7568(21)00206-3

574 4 WSIC Data Specification, v11.
575 https://www.registerfordiscover.org.uk/uploads/files/1539001703datadictionary.pdf

576 5 CPRD Aurum Data Specification, v2.8. 2022.https://cprd.com/sites/default/files/2022-
577 08/CPRD%20Aurum%20Data%20Specification%20v2.8.pdf

578 6 VA Family of EHR Cohorts (VACo Family). https://medicine.yale.edu/intmed/vacs/ (accessed 31
579 Mar 2023).

580 7 Wood A, Denholm R, Hollings S, *et al.* Linked electronic health records for research on a
581 nationwide cohort of more than 54 million people in England: data resource. *BMJ* 2021;**373**:n826.
582 doi:10.1136/bmj.n826

583 8 Data. SAIL Databank. https://saildatabank.com/data/ (accessed 31 Mar 2023).

584 9 Common Data Model (CDM) Specification, Version 6.0. 2020.https://pcornet.org/wp-
585 content/uploads/2022/01/PCORnet-Common-Data-Model-v60-2020_10_221.pdf (accessed 31
586 Mar 2023).

587 10 Wolf A, Dedman D, Campbell J, *et al.* Data resource profile: Clinical Practice Research Datalink
588 (CPRD) Aurum. *Int J Epidemiol* 2019;**48**:1740–1740g. doi:10.1093/ije/dyz034

589 11 Healthcare Epidemiology. Duke Dep. Med. https://medicine.duke.edu/education-and-
590 training/fellowship-programs/infectious-diseases/training-and-curriculum/research-
591 pathways/healthcare-epidemiology (accessed 1 Apr 2023).

592 12 Watson J, Nicholson BD, Hamilton W, *et al.* Identifying clinical features in primary care electronic
593 health record studies: methods for codelist development. *BMJ Open* 2017;**7**:e019637.
594 doi:10.1136/bmjopen-2017-019637

595 13 Sydes MR, Barbachano Y, Bowman L, *et al.* Realising the full potential of data-enabled trials in
596 the UK: a call for action. *BMJ Open* 2021;**11**:e043906. doi:10.1136/bmjopen-2020-043906

597 14 Benchimol EI, Smeeth L, Guttmann A, *et al.* The REporting of studies Conducted using
598 Observational Routinely-collected health Data (RECORD) Statement. *PLOS Med*
599 2015;**12**:e1001885. doi:10.1371/journal.pmed.1001885

600 15 Abbasizanjani H, Torabi F, Bedston S, *et al.* Harmonising electronic health records for
601 reproducible research: challenges, solutions and recommendations from a UK-wide COVID-19
602 research collaboration. *BMC Med Inform Decis Mak* 2023;**23**:8. doi:10.1186/s12911-022-02093-0

603 16 Jayatunga W, Stone P, Aldridge RW, *et al.* Code sets for respiratory symptoms in electronic
604 health records research: a systematic review protocol. *BMJ Open* 2019;**9**:e025965.
605 doi:10.1136/bmjopen-2018-025965

606 17 Pye SR, Sheppard T, Joseph RM, *et al.* Assumptions made when preparing drug exposure data
607 for analysis have an impact on results: An unreported step in pharmacoepidemiology studies.
608 *Pharmacoepidemiol Drug Saf* 2018;**27**:781–8. doi:10.1002/pds.4440

609 18 Denaxas S, Direk K, Gonzalez-Izquierdo A, *et al.* Methods for enhancing the reproducibility of
610 biomedical research findings using electronic health records. *BioData Min* 2017;**10**:31.
611 doi:10.1186/s13040-017-0151-7

612 19 VANDF (National Drug File). UMLS Metathesaurus.
613 https://www.nlm.nih.gov/research/umls/sourcereleasedocs/current/VANDF/index.html (accessed
614 1 Feb 2023).

615 20 ATC (Anatomical Therapeutic Chemical Classification System). UMLS Metathesaurus.
616 https://www.nlm.nih.gov/research/umls/sourcereleasedocs/current/ATC/index.html (accessed 1
617 Feb 2023).

618 21 RxNorm Technical Documentation.
619 2021.https://www.nlm.nih.gov/research/umls/rxnorm/docs/techdoc.html#s1_0

620 22 SNOMED International. 2.2. Value Set. In: *Practical Guide to Reference Sets*. International
621 Health Terminology Standards Development Organisation 2023.
622 https://confluence.ihtsdotools.org/display/DOCRFSPG/2.2.+Value+Set (accessed 18 Mar 2023).

623 23 Bennett Institute for Applied Data Science, University of Oxford. All BNF sections.
624 OpenPrescribing. 2023.https://openprescribing.net/bnf/ (accessed 9 Feb 2023).

24 Elkheder M, Gonzalez-Izquierdo A, Qummer Ul Arfeen M, *et al.* Translating and evaluating historic phenotyping algorithms using SNOMED CT. *J Am Med Inform Assoc JAMIA* 2023;**30**:222–32. doi:10.1093/jamia/ocac158

25 Morgan AD, Zakeri R, Quint JK. Defining the relationship between COPD and CVD: what are the implications for clinical practice? *Ther Adv Respir Dis* 2018;**12**:175346581775052. doi:10.1177/1753465817750524

26 Quint JK, Müllerova H, DiSantostefano RL, *et al.* Validation of chronic obstructive pulmonary disease recording in the Clinical Practice Research Datalink (CPRD-GOLD). *BMJ Open* 2014;**4**:e005540. doi:10.1136/bmjopen-2014-005540

27 Conway JR, Lex A, Gehlenborg N. UpSetR: an R package for the visualization of intersecting sets and their properties. *Bioinformatics* 2017;**33**:2938–40. doi:10.1093/bioinformatics/btx364

28 Lex A, Gehlenborg N, Strobelt H, *et al.* UpSet: Visualization of Intersecting Sets. *IEEE Trans Vis Comput Graph* 2014;**20**:1983–92. doi:10.1109/TVCG.2014.2346248

29 NHS Digital. NHSBSA dm+d. TRUD. https://isd.digital.nhs.uk/trud/users/guest/filters/0/categories/6/items/24/releases (accessed 1 Feb 2023).

30 Davé S, Petersen I. Creating medical and drug code lists to identify cases in primary care databases. *Pharmacoepidemiol Drug Saf* 2009;**18**:704–7. doi:10.1002/pds.1770

31 Whittaker HR, Wing K, Douglas I, *et al.* Inhaled Corticosteroid Withdrawal and Change in Lung Function in Primary Care Patients with Chronic Obstructive Pulmonary Disease in England. *Ann Am Thorac Soc* 2022;**19**:1834–41. doi:10.1513/AnnalsATS.202111-1238OC